DATA SCIENCE | DIGITAL SKILL FAIR 39.0

EXPLORATORY DATA ANALYSIS

ADVERTISING ANALYSIS

By: Siti Jahrona Rambe

› Start Slide

Intoduction

HALLO
SAYA ONA

Seorang mahasiswa semester 6 di Universitas Sumatera Utara yang memiliki ketertarikan dalam bidang Data Analyst. Dalam proyek ini, saya melakukan eksplorasi data pada dataset Advertising untuk memahami pengaruh pengeluaran iklan di TV, Radio, dan Newspaper terhadap penjualan (Sales).

DATA OVERVIEW

Data yang digunakan adalah data Advertising yang berasal dari kaggle dimana untuk memahami pengeluaran iklan di TV, Radio, dan Newspaper terhadap penjualan (Sales).

Analisis yang dilakukan adalah:

1. Importing data.
2. Preprocessing data (missing value, duplikat, outlier).
3. Analisis data & Visualisasi Data.

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	NaN	58.5	18.5
4	180.8	NaN	58.4	12.9

ABOUT DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   TV         200 non-null    float64
 1   Radio      188 non-null    float64
 2   Newspaper  198 non-null    float64
 3   Sales      200 non-null    float64
dtypes: float64(4)
memory usage: 6.4 KB
```

Dataset ini berisi 200 data observasi dengan 4 fitur numerik yang merepresentasikan pengeluaran iklan dan penjualan produk:

- **TV, Radio, dan Newspaper:** Besarnya anggaran iklan di masing-masing media.
- **Sales:** Jumlah produk yang terjual.

1. Terdapat 12 nilai kosong pada kolom Radio dan 2 nilai kosong pada kolom Newspaper.
2. Terdapat 6 baris data yang terduplikasi
3. Ditemukan keberadaan outlier pada kolom Newspaper.

01 Import Data

Import data dengan pandas from google drive

```
[3]  #syntax menggunakan Pandas
     import pandas as pd
```

```
[4]  from google.colab import drive
     drive.mount('/content/gdrive/', force_remount=True)
```

Kemudian untuk membaca file CSV dengan pemisah titik koma (;). Tanpa delimiter=';', pandas akan membaca seluruh baris sebagai satu kolom karena tidak menemukan pemisah koma seperti default-nya.

```
# Tanpa delimiter=';', pandas akan mengira itu semua hanya 1 kolom karena tidak ada koma.
df = pd.read_csv('/content/gdrive/MyDrive/advertising2.csv' , delimiter=';')
df
```

02 Preprocessing Data

Data Problem

1. Terdapat 12 nilai kosong pada kolom Radio dan 2 nilai kosong pada kolom Newspaper.

```
# cari nilai yang kosong
df.isnull().sum()
```

	0
TV	0
Radio	12
Newspaper	2
Sales	0

dtype: int64

2. Terdapat 6 baris data yang terduplikasi

```
[43] df.duplicated().sum()
```

```
np.int64(6)
```

3. Ditemukan keberadaan outlier pada kolom Newspaper.

```
cek_outlier_iqr(df, 'TV')
cek_outlier_iqr(df, 'Radio')
cek_outlier_iqr(df, 'Newspaper')
```

```
📌 Kolom: TV
Jumlah outlier: 0

📌 Kolom: Radio
Jumlah outlier: 0

📌 Kolom: Newspaper
Jumlah outlier: 2
```

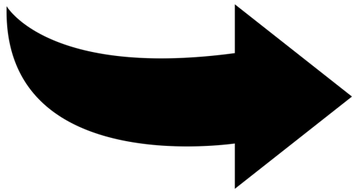
02 Preprocessing Data

1. Handling Missing Values

Karena jumlah data kosong tidak terlalu banyak dibandingkan total data, maka bisa memilih imputasi (isi nilai kosong) agar data tidak banyak terbuang.

Karena kolom bertipe numerik maka isi dengan mean

```
# isi missing values dengan nilai mean
df['Radio'] = df['Radio'].fillna(df['Radio'].mean())
df['Newspaper'] = df['Newspaper'].fillna(df['Newspaper'].mean())
```



Data informasi setelah dilakukan handling missing values

```
df.isnull().sum()
```

	0
TV	0
Radio	0
Newspaper	0
Sales	0

dtype: int64

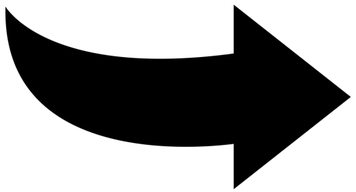
Tidak ada lagi missing values

02 Preprocessing Data

2. Handling Duplicate Data

Menghapus data yang duplikat

```
# Hapus data yang duplikat
df = df.drop_duplicates()
```



Data informasi setelah menghapus data duplikat

```
# Cek data yang duplikat lagi
df.duplicated().sum()
```

```
np.int64(0)
```

Tidak ada lagi data yang duplikat

02 Preprocessing Data

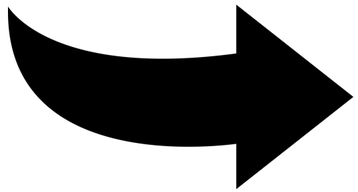
3. Handling Outlier

Ganti nilai outlier dengan median jika tidak ingin datanya hilang

```
median_value = df['Newspaper'].median()
df.loc[[16, 101], 'Newspaper'] = median_value
```

Kemudian cek kembali outlier dengan IQR

```
def cek_outlier_iqr(df, kolom):
    Q1 = df[kolom].quantile(0.25)
    Q3 = df[kolom].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = df[(df[kolom] < lower_bound) | (df[kolom] > upper_bound)]

    print(f"\n📌 Kolom: {kolom}")
    print(f"Jumlah outlier: {len(outliers)}")
    return outliers
```



Data informasi setelah handling outlier

```
cek_outlier_iqr(df, 'TV')
cek_outlier_iqr(df, 'Radio')
cek_outlier_iqr(df, 'Newspaper')
```

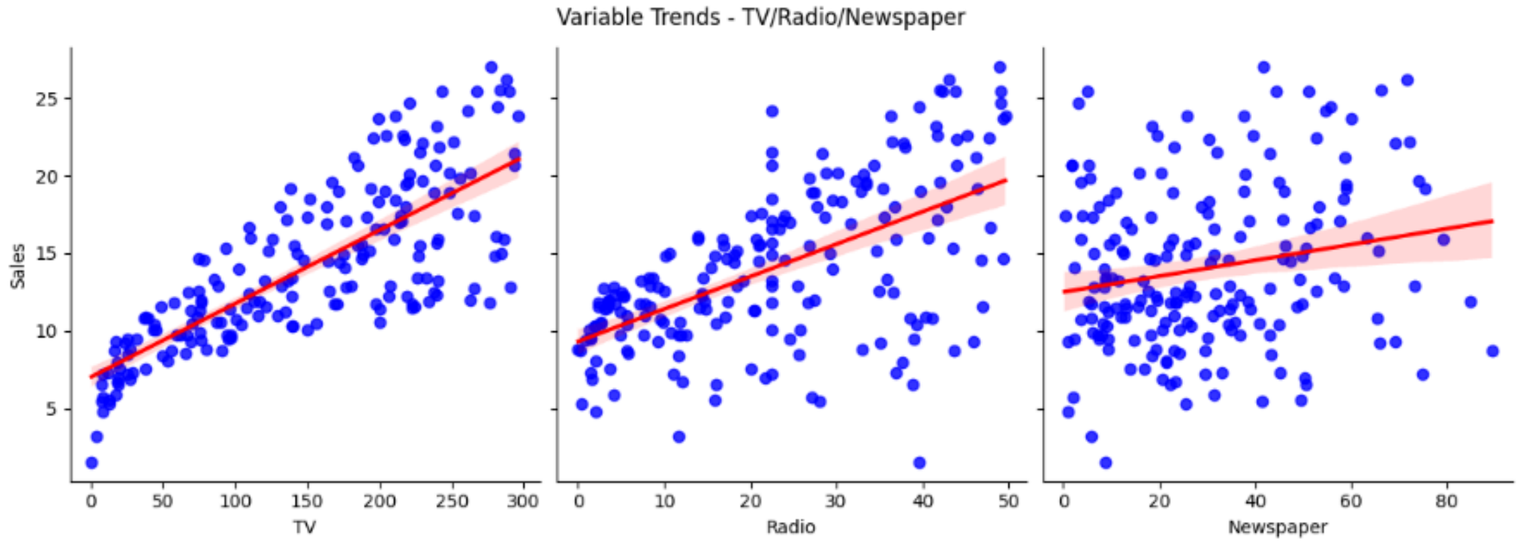
```
📌 Kolom: TV
Jumlah outlier: 0

📌 Kolom: Radio
Jumlah outlier: 0

📌 Kolom: Newspaper
Jumlah outlier: 0
```

Tidak ada lagi data yang outlier

03 Analisis data & Visualisasi Data

Korelasi Variabel dengan Scatter Plot dengan Garis Regresi

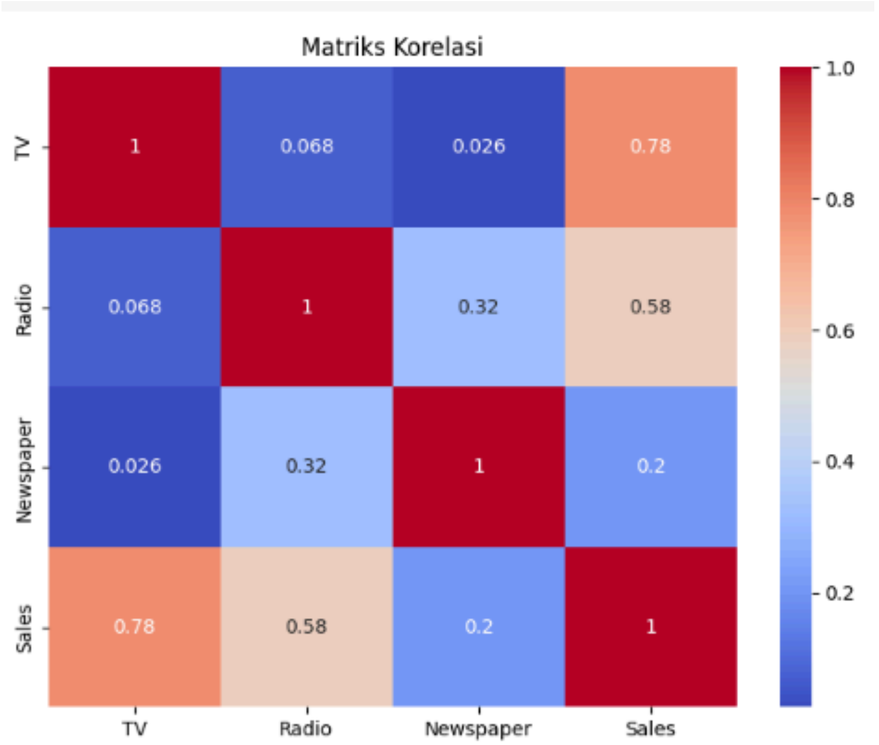
Visualisasi ini menunjukkan hubungan antara jumlah pengeluaran iklan di TV, Radio, dan Newspaper terhadap penjualan (Sales). Titik-titik biru merepresentasikan data, sementara garis merah menunjukkan tren regresi linier. Dari ketiga grafik, terlihat bahwa **iklan TV memiliki pola hubungan paling kuat dan konsisten terhadap peningkatan penjualan,** diikuti oleh Radio. Sementara Newspaper menunjukkan sebaran data yang lebih acak dan hubungan yang lebih lemah terhadap penjualan.



03 Analisis data & Visualisasi Data

Korelasi Variabel Heatmap

Heatmap ini menggambarkan kekuatan hubungan antar variabel dengan nilai korelasi. Hasilnya menunjukkan bahwa **TV memiliki korelasi paling tinggi terhadap Sales (0.78),** diikuti oleh Radio (0.58), dan Newspaper hanya (0.20). Ini mengindikasikan bahwa strategi pemasaran melalui TV lebih efektif dalam mendorong penjualan dibandingkan dengan media lainnya.

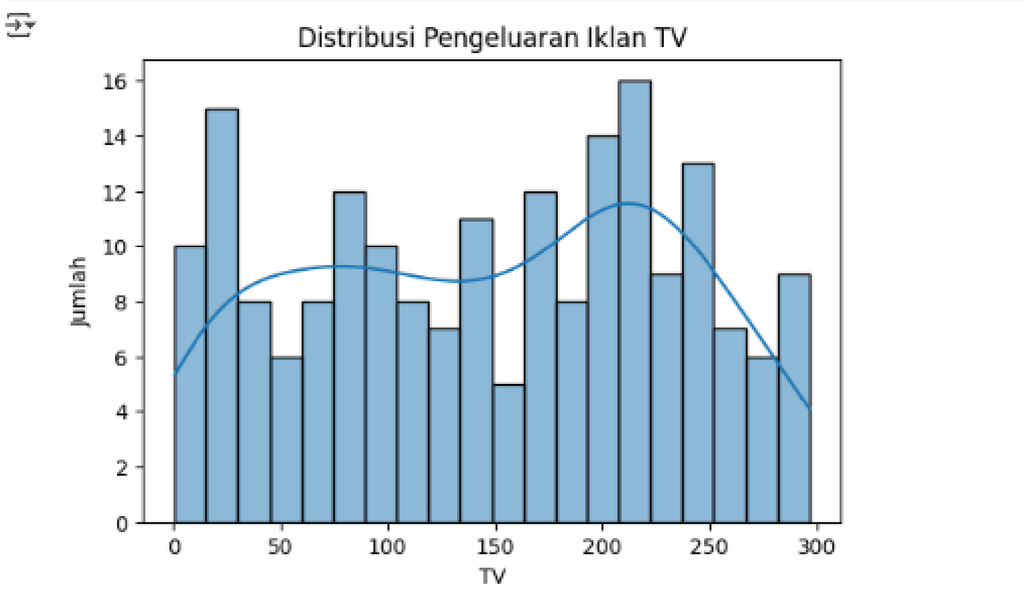


03 Analisis data & Visualisasi Data

Distribusi Pengeluaran Iklan TV

Distribusi iklan TV terlihat cukup merata dengan sedikit kecenderungan ke tengah (sekitar 150–250). Ini menunjukkan bahwa banyak perusahaan mengalokasikan dana iklan yang cukup besar untuk TV, dan distribusinya menyebar luas.

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
sns.histplot(data=df, x='TV', bins=20, kde=True)
plt.title('Distribusi Pengeluaran Iklan TV')
plt.xlabel('TV')
plt.ylabel('Jumlah')
plt.show()
```

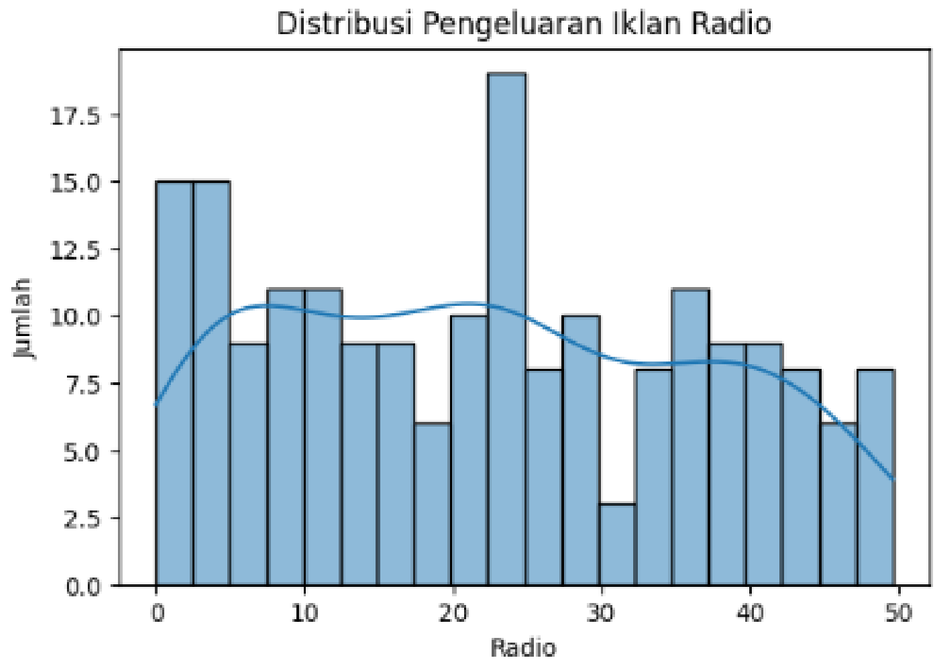


03 Analisis data & Visualisasi Data

Distribusi Pengeluaran Iklan Radio

Distribusi iklan radio tampak cenderung menyebar rata di berbagai tingkat pengeluaran, namun lebih banyak terjadi pada kisaran 10–40. Ini menunjukkan bahwa pengeluaran iklan radio cukup bervariasi, meski tidak sebesar TV.

```
plt.figure(figsize=(6,4))
sns.histplot(data=df, x='Radio', bins=20, kde=True)
plt.title('Distribusi Pengeluaran Iklan Radio')
plt.xlabel('Radio')
plt.ylabel('Jumlah')
plt.show()
```

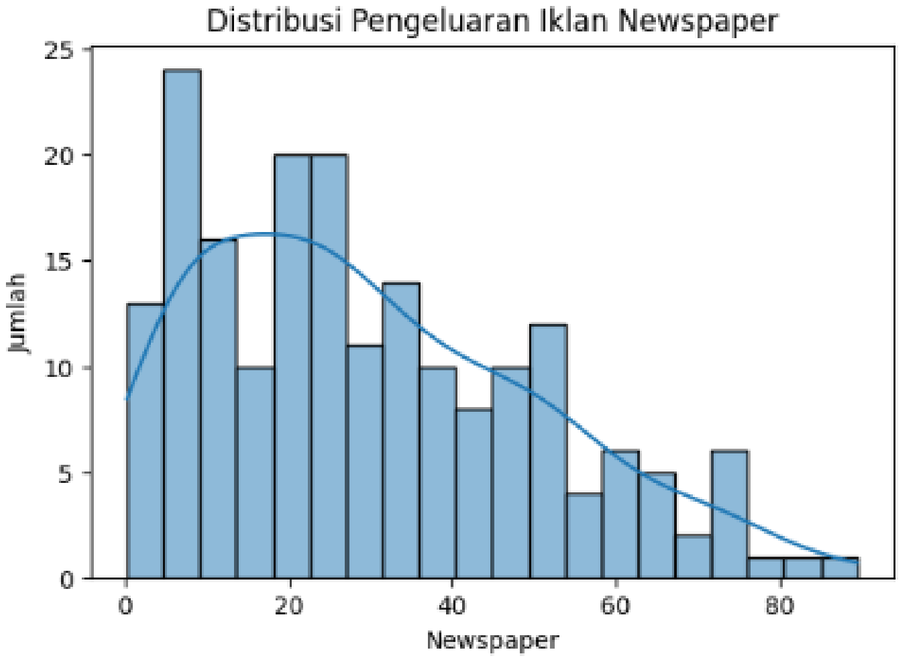


03 Analisis data & Visualisasi Data

Distribusi Pengeluaran Iklan Newspaper

Distribusi iklan koran menunjukkan pola yang menurun. Sebagian besar pengeluaran ada di bawah 30, dan sangat sedikit yang mengalokasikan lebih dari 60. Ini menandakan bahwa koran bukanlah media utama untuk iklan dibanding TV atau radio.

```
plt.figure(figsize=(6,4))
sns.histplot(data=df, x='Newspaper', bins=20, kde=True)
plt.title('Distribusi Pengeluaran Iklan Newspaper')
plt.xlabel('Newspaper')
plt.ylabel('Jumlah')
plt.show()
```



THANK YOU

By: Siti Jahrona Rambe

› End Slide